Ульяновский государственный университет
Факультет математики, информационных и авиационных технологий
Кафедра математического моделирования технических систем

Павлов П.Ю.

Большие данные и методы машинного обучения в исследованиях

Методические указания
для самостоятельной работы аспирантов

Ульяновск, 2022

Печатается по решению Ученого совета ФМИАТ
Ульяновского государственного университета

Павлов П.Ю.
Большие данные и методы машинного обучения в исследованиях**:**
Методические указания для самостоятельной работы аспирантов **/** П.Ю. Павлов. – Ульяновск: УлГУ, 2022. – 13 с.

Методическое пособие по дисциплине «Большие данные и методы машинного обучения в исследованиях» предназначено в помощь аспирантам, обучающимся по всем научным специальностям, для самостоятельного изучения отдельных разделов курса. Методические указания включают в себя требования к результатам освоения дисциплины, тематический план дисциплины, список рекомендуемой литературы, контрольные вопросы к зачету.

СОДЕРЖАНИЕ

1. Цели и задачи освоения дисциплины

Цель освоения дисциплины «Большие данные и методы машинного обучения в исследованиях» состоит в формировании у аспирантов:

- знаний наиболее актуальных работ в области применения новых типов данных в разных направлениях науки;
- навыков по сбору данных из социальных медиа и других цифровых следов с использованием языка программирования Python;
- навыков обработки и анализа различных типов данных (сетевые, текстовые и геоданные) с использованием языка программирования Python.

2. Перечень планируемых результатов освоения дисциплины (модуля)

В результате изучения дисциплины аспирант должен:

Знать:

- основные теоретические, методологические и практические подходы к анализу больших данных и новых типов данных;
- источники новых типов данных;
- ключевые исследовательские работы и направления в области применения больших данных и методов машинного обучения в естественнонаучных, медико-биологических и общественно-гуманитарных науках;
- основные этические принципы работы с данными и этические проблемы, связанные с использованием больших данных;

Уметь:

- ставить исследовательские вопросы и формулировать гипотезы, протестировать которые можно с использованием больших данных;
- грамотно использовать алгоритмы машинного обучения и статистического анализа для изучения больших данных;
- проводить исследование полного цикла с использованием новых типов данных;

- интерпретировать и оформлять полученные результаты.

Иметь навыки:

- Постановки исследовательского вопроса в области применения больших данных, данных нового типа и методов машинного обучения в исследованиях образования;
- планирование исследования полного цикла;
- презентации и защиты индивидуального исследовательского проекта.

3. Список рекомендуемой литературы для самостоятельной работы аспирантов

а) Список рекомендуемой литературы
Основная:

Федоров, Д. Ю. Программирование на языке высокого уровня Python : учебное пособие для вузов / Д. Ю. Федоров. — 4-е изд., перераб. и доп. — Москва : Издательство Юрайт, 2022. — 214 с. — (Высшее образование). — ISBN 978-5-534-15733-8. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/509562

Платонов, А. В. Машинное обучение : учебное пособие для вузов / А. В. Платонов. — Москва : Издательство Юрайт, 2023. — 85 с. — (Высшее образование). — ISBN 978-5-534-15561-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/520544

Черткова, Е. А. Статистика. Автоматизация обработки информации : учебное пособие для вузов / Е. А. Черткова. — 2-е изд., испр. и доп. — Москва : Издательство Юрайт, 2023. — 195 с. — (Высшее образование). — ISBN 978-5-534-01429-7. — Текст : электронный // Образовательная платформа Юрайт [сайт]. — URL: https://urait.ru/bcode/513393

Дополнительная:

Маккинли У., Python и анализ данных [Электронный ресурс] / Уэс Маккинли - М. : ДМК Пресс, 2015. - 482 с. - ISBN 978-5-97060-315-4 - Режим доступа: http://www.studentlibrary.ru/book/ISBN9785970603154.html

Рощин, С. М. Современные интернет-технологии. Семь главных трендов : научно-популярное издание / С. М. Рощин. - 2-е изд. - Москва : Дашков и К, 2022. - 124 с. - ISBN 978-5-394-04846-3. - Текст : электронный. - URL: https://znanium.com/catalog/product/1927306

Цифровой бизнес : учебник / под науч. ред. О.В. Китовой. — Москва : ИНФРА-М, 2023. — 418 с. — (Высшее образование). — DOI 10.12737/textbook_5a0a8c777462e8.90172645. - ISBN 978-5-16-013017-0. - Текст : электронный. - URL: https://znanium.com/catalog/product/1917620

Статистическая обработка экспериментальных данных. Регрессионный анализ в языке R : учебное пособие / В. Ю. Потапова, А. С. Тарасов, Е. С. Геращенко, М. Б. Никифоров. — Рязань : РГРТУ, 2018. — 52 с. — ISBN 978-5-6041320-7-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/168238

Согласовано:

ДИРЕКТОР НБ / БУРХАНОВА М.М. / подпись / 15.04.2022
Должность сотрудника НБ — ФИО — подпись — дата

4. Разделы дисциплины и виды учебных занятий

Название и разделов, и тем	Всего	Виды учебных занятий		
		Аудиторные занятия		Самостоятельная работа
		лекции	практические занятия, семинар	
1	2	3	4	5
Тема 1. Большие	13	2	2	9

данные и методы машинного обучения в естественнонаучных, медико-биологических и общественно-гуманитарных науках				
Тема 2. Введение в язык программирования python.	13	2	2	9
Тема 3. Автоматический сбор данных из интернета	13	2	2	9
Тема 4. Использование методов машинного обучения для предсказания характеристик пользователей на основании их цифровых следов	13	2	2	9
Тема 5. Интеллектуальный анализ текстов	14	2	2	10
Тема 6. Анализ геопространственных данных	14	2	2	10
Тема 7. Этика использования больших данных	14	2	2	10
Тема 8. Презентация индивидуального исследовательского проекта	14	2	2	10
Итого	**108**	**16**	**16**	**76**

Изучение дисциплины предусматривает 16 часов лекционных и 16 часов семинарских занятий. 76 часов отведено на самостоятельное изучение дисциплины.

5. Тематический план дисциплины

Тема 1. Большие данные и методы машинного обучения в естественнонаучных, медико-биологических и общественно-гуманитарных науках.

Новые типы данных: интернет-данные, другие цифровые следы и возможности их применения. Обсуждение идей индивидуальных исследовательских проектов.

Тема 2. Введение в язык программирования python.

Базовые типы данных. Переменные. Операторы. Условия, циклы и функции. Ошибки и предупреждения.

Тема 3. Автоматический сбор данных из интернета.

Форматы данных. HTML и JSON. Использование API интернет-сервисов на примере социальной сети ВКонтакте. Анализ социальных сетей: основные теоретические понятия и приложения. Изучение сетей дружбы на примере данных «ВКонтакте».

Тема 4. Использование методов машинного обучения для предсказания характеристик пользователей на основании их цифровых следов.

Анализ последовательностей. Прогнозирование и визуализация данных

Тема 5. Интеллектуальный анализ текстов.

Основные теоретические понятия и приложения. Тематическое моделирование. Анализ текстов из социальных сетей.

Тема 6. Анализ геопространственных данных.

Основные теоретические понятия и приложения. Методы сбора, практическое использование и интерпретация результатов.

Тема 7. Этика использования больших данных.

Алгоритмы и дискриминация. Применение технологий больших данных для задач управления в банковской, страховой, финансовой индустриях.

Тема 8. Презентация индивидуального исследовательского проекта.

6. Тематика семинарских и практических занятий

Тема 1. Большие данные и методы машинного обучения в естественнонаучных, медико-биологических и общественно-гуманитарных науках.

Вопросы для дискуссии:

1. Новые типы данных: интернет-данные, другие цифровые следы и возможности их применения.
2. Обсуждение идей индивидуальных исследовательских проектов.

Тема 2. Введение в язык программирования python.

Вопросы для дискуссии:

1. Базовые типы данных. Переменные. Операторы.

2. Условия, циклы и функции. Ошибки и предупреждения.

Тема 3. Автоматический сбор данных из интернета.

Вопросы для дискуссии:

1. Форматы данных. HTML и JSON. Использование API интернет-сервисов на примере социальной сети ВКонтакте.

2. Анализ социальных сетей: основные теоретические понятия и приложения. Изучение сетей дружбы на примере данных «ВКонтакте».

Тема 4. Использование методов машинного обучения для предсказания характеристик пользователей на основании их цифровых следов.

Вопросы для дискуссии:

1. Анализ последовательностей.

2. Прогнозирование и визуализация данных.

Тема 5. Интеллектуальный анализ текстов.

Вопросы для дискуссии:

1. Основные теоретические понятия и приложения.
2. Тематическое моделирование. Анализ текстов из социальных сетей.

Тема 6. Анализ геопространственных данных.

Вопросы для дискуссии:

1. Основные теоретические понятия и приложения.
2. Методы сбора, практическое использование и интерпретация результатов.

Тема 7. Этика использования больших данных.

Вопросы для дискуссии:

1. Алгоритмы и дискриминация.
2. Применение технологий больших данных для задач управления в банковской, страховой, финансовой индустриях.

Тема 8. Презентация индивидуального исследовательского проекта.

7. Контрольные вопросы по дисциплине (вопросы к зачету)

1. Понятие Большие данные. Роль цифровой информации в 21 веке.
2. Виды массивов данных.
3. Базовые принципы обработки больших данных.
4. Технологии обработки больших данных: NoSQL, MapReduce, Hadoop, R.
5. Технологии Business Intelligence и реляционные системы управления базами данных.
6. Прогнозирование и предвидение: общее и особенное.
7. Виды прогнозов.

8. Вопросы безопасности больших данных.

9. Основные описательные статистики.

10. Регрессионный анализ.

11. Основная идея дисперсионного анализа.

12. Сущность кластерного анализа.

13. Дискриминантный анализ: модель и общая процедура выполнения.

14. Цели факторного анализа.

15. Программные средства анализа данных: Statistica, SPSS, Excel; их преимущества и недостатки.

8. Методические указания к семинарским занятиям

Семинар – форма занятия, обеспечивающая создание аспирантами личных образовательных продуктов в ходе коллективно-групповой коммуникации. Семинары отличаются от других видов занятий повышенной активностью и самостоятельностью обучающихся, возможностью проявления их способностей к организации деятельности. По способу и характеру проведения различают вводные, обзорные, самоорганизующие, поисковые, индивилуальные и групповые семинары, семинары проекты, семинары по решению задач, круглые столы, «мозговые штурмы», семинары деловые игры и др. Для эффективного участия в семинаре подготовку к нему рекомендуется вести в следующей последовательности.

• необходимо ознакомиться с содержанием очередной темы по программе;

• используя рекомендуемую литературу и конспекты лекций, следует изучить все положения данной темы и ответить на проблемные вопросы;

• при изучении литературы необходимо делать краткие выписки, что способствует лучшему усвоению материала и облегчает его использование в ходе семинара.

Требования к качеству подготовки аспирантов к практическим (семинарским) занятиям:

1. Подготовка к практическим (семинарским) занятиям является обязательной частью работы аспиранта и производится по всем вопросам темы, указанным в плане занятия, а не выборочно по отдельным вопросам. Сплошная подготовка способствует полноценному освоению темы и эффективной работе практического (семинарского) занятия.

2. Работа аспиранта на практическом (семинарском) занятии предполагает его высокую активность и соответствие следующим требованиям при публичном выступлении:

а) свободное устное воспроизведение подготовленного выступления по вопросам с использованием мини-конспектов в качестве вспомогательного средства;

б) готовность и умение отвечать на вопросы и делать выводы из сказанного;

в) владение терминологией курса «Большие данные и методы машинного обучения в исследованиях»;

г) временной регламент выступления 7-10 минут.

3. После завершения изучения курса аспирант должен владеть основными концепциями курса и использовать их для обсуждения вопросов дисциплины «Большие данные и методы машинного обучения в исследованиях».

9. Методические рекомендации по организации самостоятельной работы аспирантов

Внеаудиторная самостоятельная работа аспирантов (далее СРА) – планируемая учебная, учебно-исследовательская, научно-исследовательская работа аспирантов, выполняемая во внеаудиторное время по заданию и при методическом руководстве преподавателя, но без его непосредственного участия.

Цель СРА - осмысленно и самостоятельно работать сначала с учебным материалом, затем с научной информацией, развивать основы самоорганизации и самовоспитания с тем, чтобы привить умение в дальнейшем непрерывно повышать свою квалификацию.

Целью СРА по дисциплине «Большие данные и методы машинного обучения в исследованиях» является овладение фундаментальными знаниями, профессиональными умениями и навыками решения задач и теоретическим материалом по данной дисциплине. СРА способствует развитию самостоятельности, ответственности и организованности, творческого подхода к решению различных проблем. Объем СРА определяется учебным планом программы аспирантуры. СРА является обязательной для каждого аспиранта.

Для успешной организации СРА необходимы следующие условия:

- готовность аспирантов к самостоятельной работе по данной дисциплине и высокая мотивация к получению знаний;

- наличие и доступность необходимого учебно-методического и справочного материала;

- регулярный контроль качества выполненной самостоятельной работы (проверяет преподаватель на коллоквиумах);

- консультационная помощь преподавателя.

При изучении дисциплины организация СРА должна представлять единство трех взаимосвязанных форм:

1. Внеаудиторная самостоятельная работа;

2. Аудиторная самостоятельная работа, которая осуществляется под непосредственным руководством преподавателя;

3. Творческая, в том числе научно-исследовательская работа.